Building a Library of Eligibility Criteria to Support Design of Clinical Trials

Krystyna Milian[1,2], Anca Bucur[2], and Frank van Harmelen[2]

[1] Vrije Universiteit, Amsterdam
[2] Philips Research, Eindhoven
k.milian@vu.nl, anca.bucur@philips.com, frank.van.harmelen@cs.vu.nl

Abstract. The completion of clinical trial depends on sufficient participant enrollment, which is often problematic due to the restrictiveness of eligibility criteria, and effort required to verify patient eligibility. The objective of this research is to support the design of eligibility criteria, enable the reuse of structured criteria and to provide meaningful suggestions of relaxing them based on previous trials. The paper presents the first steps, a method for automatic comparison of criteria content and the library of structured and ordered eligibility criteria that can be browsed with the fine-grained queries. The structured representation consists of the automatically identified contextual patterns and semantic entities. The comparison of criteria is based on predefined relations between the patterns, concept equivalences defined in medical ontologies, and finally on threshold values. The results are discussed from the perspective of the scope of the eligibility criteria covered by our library.

Keywords: Modeling clinical trial data, Semantic annotation, Medical ontologies, Eligibility criteria, Formalization of eligibility criteria, Supporting design of eligibility criteria.

1 Introduction

Insufficient recruitment is often a barrier that prevents the finalization of a clinical trial and obtaining evidence about new prevention, diagnostic or treatment methods. The recruitment process is time and effort consuming, as for each patient that is considered for enrollment it requires verification of whether the patient satisfies all eligibility criteria of the trial. Additionally, the completion of the trial depends on enrolling a sufficient number of participants. Applications assisting the investigators while designing a trial and further when recruiting patients could help to automate the process.

A few studies have addressed the task of supporting the verification of patient eligibility [1]. However, little attention has been devoted to the support of the design of eligibility criteria. The main purpose of the study reported here was to address this issue. Our objective is to enable the reuse of structured eligibility criteria of existing trials, which can be used to provide meaningful suggestions to trial designers during the definition of new sets of criteria, for example concerning the revision of unnecessarily restrictive conditions.

A. ten Teije et al. (Eds.): EKAW 2012, LNAI 7603, pp. 327–336, 2012.

We approached the problem by analyzing eligibility criteria of studies published at ClinicalTrials.gov. We extended our previous work on formalization of criteria using contextual patterns [2] by enlarging the set of patterns and identifying in criteria semantic entities i.e. ontology concepts, measurements and numbers. Next, we applied it to automatically structure and classify eligibility criteria of breast cancer trials. Further, we designed a method for comparing the criteria content and their restrictiveness. Using obtained results we created a library of structured eligibility criteria. The paper contains examples describing its content and possible usage. In our future work we will connect the formalized eligibility criteria with queries, which will be used to assess whether a given patient satisfies the entry conditions of a trial. Since many eligibility criteria are very similar or even identical across the trials, reusing computable criteria could significantly enhance the recruitment process.

The paper is organized as follows. The next section introduces a method we developed to interpret eligibility criteria by first formalizing the meaning of the criteria and then comparing their restrictiveness. Section 3 describes the model of the library and the way it was populated with data: formalized criteria of existing trials. Further, section 4 gives quantified results of the library content. Related work is described in section 5, the last chapter contains conclusions.

2 Interpreting Eligibility Criteria

This section describes our method designed to build a library of structured eligibility criteria. Our aim was to enable the reuse of structured representation, and provide meaningful suggestions of relaxing criteria to enable enrolling a larger number of participants. Because of similarities and repetitions of criteria across the trials, our claim is that by formalizing eligibility criteria of a large corpus of clinical trials we can create a sufficiently rich library to serve the task. Our method relies on:

1. Extracting eligibility criteria from a corpus of publicly available clinical trials
2. Formalizing their content
3. Identifying similarities between the criteria and determining relations, i.e. which one is more restrictive

2.1 The Method for Formalizing Eligibility Criteria

The method of formalization of eligibility criteria consists of several steps, depicted in Figure 1.

We start with the pre-processing of criteria, delimiting the sentences using GATE [3], the open source framework for text processing. Next, whenever possible, we recognize the domain of the criteria, e.g. Age, Cardiovascular, Chemotherapy etc. Further follow the two main steps of criteria formalization.

First, we recognize the general meaning of a criterion, by detecting the patterns providing the contextual information about the semantic entities mentioned in the criterion. The set of patterns was initially described in our previous

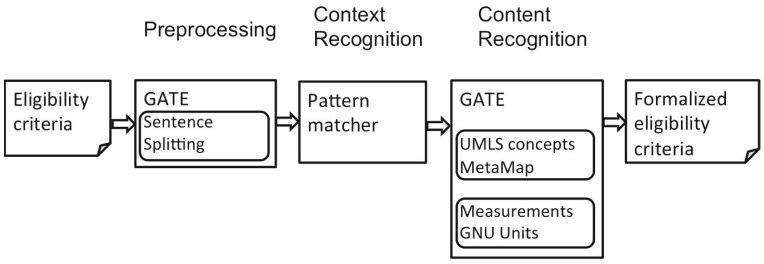

Fig. 1. The pipeline of processing steps of eligibility criteria

work [2] and further extended. It was manually defined by analyzing eligibility criteria published at ClinicalTrials.gov and contains 165 items that reflect the typically occurring constraints. The patterns cover criteria related to patient characteristics (e.g. Age over ()), disease characteristics (e.g. T () stage) and prior and concurrent therapies (e.g. No concurrent () except for ()). They are classified according to several dimensions that characterize the content of corresponding eligibility criteria from various perspectives:

- Temporal status (TS): prior, current, planned event
- Time independent status (TIS): present, absent, conditional
- Constraint types (CT): temporal (start, end, duration), confirmation, co-occurrence, exception, inclusion
- Subject: patient, family of a patient

The algorithm of pattern identification is based on regular expressions, it finds the longest patterns together with the nested ones. In total we defined 468 regular expressions corresponding to the 165 patterns.

Detection of patterns enables recognizing the context in which semantic entities occur. Next, we identify these semantic entities, which can be instantiated by diseases, treatments, lab measurement, value or temporal constraints. We approached the task by incorporating state of the art tools i.e. GATE the NLP framework, providing a library of semantic taggers, and MetaMap, an UMLS [4] ontology annotator. In the workflow of text processing steps we used a tokenizer, sentence splitter, Number, Measurement and MetaMap taggers, wrapped in our application using the GATE API. A result of MetaMap annotation is metadata about identified mapping (or a list of candidates), the UMLS concept id, its preferred name, semantic type (ST), score of mapping, and list ontologies covered by UMLS that specify the concept. The measurement plugin, based on GNU Units [5], recognizes the measurements, including value, unit and dimension, and additionally normalizes the values according to the standard units. Recognition of mentioned entities enables the interpretation of criteria meaning and processing of normalized representation (terms identified in text can be replaced by unique UMLS identifiers, measurements by normalized values and units).

Following example illustrates the approach. In criterion: 'No prior malignancy except for nonmelanoma skin cancer', first, we detect the pattern 'No prior ()

for ()', and second, the concepts 'malignancy' and 'nonmelanoma skin cancer'. To evaluate criteria, the patterns can be linked to predefined templates, the incomplete queries, which after filling with the semantic entities identified and mapped to corresponding items in patient record, can be executed to verify patient eligibility.

2.2 The Method for Comparing Eligibility Criteria

By formalizing eligibility criteria we have created the basis for automated mining of the criteria content. We used obtained results to determine relations between concrete eligibility criteria, in order to assist the designer with proposing alternative, less restrictive, but still meaningful suggestions. This section describes our approach to the criteria comparison based on identified context patterns, ontology concepts and value constraints.

Comparison of Criteria That Match the Same Context Pattern. Recognizing syntactic patterns enables capturing the general meaning of criteria. Information that two criteria match the same pattern provides valuable information about their similarity. Further, depending on their instantiation, they can be classified as comparable. For instance, although the two following criteria: No chemotherapy within last month and No prior lung cancer with last year match the same pattern: No prior () within (), comparing them for our purpose is irrelevant. Criteria can be compared when have the same main argument and different value constraints, i.e.:

- Lower or upper thresholds for lab values, e.g. Bilirubin less than 2.0 mg/dL can be compared with: Bilirubin less than 1.5 mg/dL.
- Temporal constraints, which restrict: start, end or duration of some medical event, for example: At least 1 week since prior hormonal therapy can be compared with At least 4 weeks since prior hormonal therapy.

In both cases the comparison is possible when the values have the same normalized unit identified by MetaMap.

Comparison Based on the Relations between the Context Patterns. To compare criteria with different syntax we designed another strategy. We predefined relations between some patterns (canRelax, canBeRelaxedBy), indicating which pattern can be relaxed by which. These relations express the possibility that corresponding criteria can be in the relation isMoreRelaxed/ isMoreStrict, when they are instantiated with the same argument. The relations between the patterns are based on:

- Explicitly stated exceptions e.g.: No prior () can be relaxed by: No prior () unless (), No prior () except for ()
- Specified value constraints: temporal, confirmation, number of occurrences. The constraints, depending on the context (Time independent status), relax or restrict the primary pattern, for example:

- No prior () can be relaxed by: No () within (), At least () since ()
- History of () within () or History of () confirmed by () can be relaxed by: History of (), because the latter requires the presence of the event at any point in time, and does not restrict the evidence type.

In total we have defined 36 relaxing relations between the patterns.

Using the described methods for the formalization and comparison of eligibility criteria, we processed inclusion and exclusion criteria from the corpus of clinical trials and populated the library.

3 Library of Eligibility Criteria

3.1 The Model of the Library

This section describes the model of the library of eligibility criteria, designed to reflect the most relevant information about their content.

The library was modeled as ontology to enhance semantic reasoning. The lists of classes, and properties are displayed in Figure 2. The model captures data related to Trial (hasID, hasCriterion), Criterion (hasContent, hasDomain), its Dimensions of classification (hasTemporalStatus, hasTimeIndependentStatus, hasSubject, etc) the formalization of its Content - one from a set of Pattern Instance or Concept. Pattern Instances have modeled corresponding value constraints (hasContent, hasValue, see the list of object and data (sub)properties). Concept has specified its metadata (hasConceptId - UMLS id, hasSemanticType, hasSource - definining ontology and occursIn - links to the criteria where it occurs). Additionally, the model explicitly defines transitive relations between the patterns (canRelax/canBeRelaxedBy), and concrete criteria (isMoreRelaxed/isMoreStrict). The criteria and extracted data are represented as individuals. Modeling the library as an ontology enables sharing it, extending or linking to other sources.

3.2 Populating the Model

Clinical trials that were used to build the library of criteria come from the ClinicalTrials.gov repository, a service of the U. S. National Institute of Health, containing data about clinical trials conducted worldwide. We focused on clinical trials related to breast cancer and processed eligibility criteria from 300 studies. The model was populated using the results of the processing steps described in the previous section. Firstly we split the sentences, next we recognized corresponding patterns and the semantic entities mentioned. For the simplification purpose we took into account only the criteria that match a single pattern.

Each pattern has labelled arguments in order to facilitate the task and correctly associate recognized items. For example a pattern 'No prior () within () except for ()' has labelled its 3 arguments as: main argument, end time constraint and exception, which after detection were saved as values of corresponding object or data properties. Finally, we compared corresponding criteria. The results were saved as triples using the OWL API.

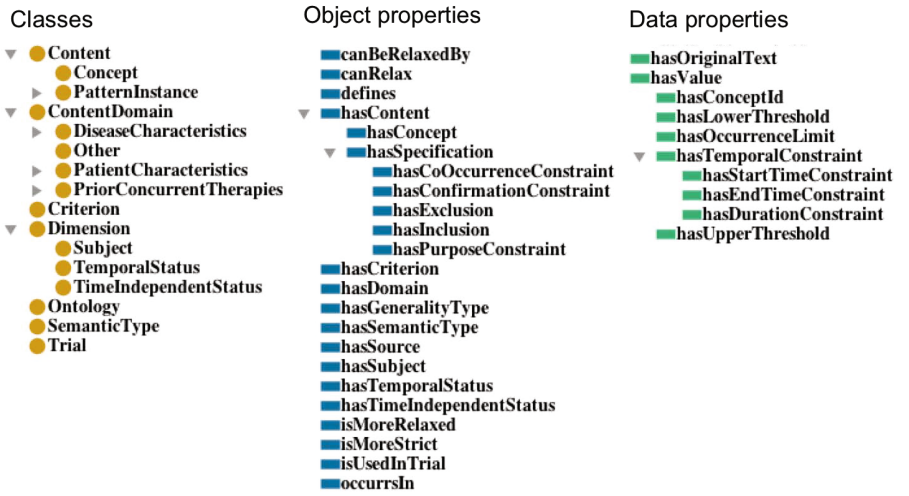

Fig. 2. Library model

4 Results

4.1 Characteristics of the Library

This section describes the final result of populating the library. Its content is quantitatively characterized in Table 4.1. The library contains 1799 structured eligibility criteria out of 10067 (17.9%) used in the experiment, which come from 268 different clinical trials out of all 300 processed. This result indicates the need of improving the recall of the method. One limitation is caused by the fact that our method takes into account only criteria that match one pattern, while many of them are more complex. The interpretation of such criteria would require correct identification of relations of recognized patterns i.e. conjunction, disjunction, nesting. Filtering out criteria, which were matched to some pattern, but the annotation of their arguments with ontology concept by MetaMap did not return any results, caused another reason of low recall.

The table contains also information about the ontology concepts identified, i.e. 1241 UMLS concepts were recognized that belong to 91 various semantic types, and are defined in 46 ontologies covered by UMLS. With respect to the result

Table 1. The library characteristics

Eligibility criteria	1799/10067 (17.9 %)
Trials	268/300 (89.3 %)
Concepts	1241
Semantic Types	91
Ontologies from UMLS	46
Relaxations based on value threshold	202
Relaxations based on semantic modifiers	87

of criteria comparison, in total the algorithm identified 289 cases of eligibility criteria that could be potentially relaxed by one of the other conditions included in the library. This accounts for 16% of the entire number of formalized criteria.

Table 2 characterizes the type of formalized criteria, by giving number of criteria belonging to a few major classes (not mutually exclusive).

Table 2. The characteristics of formalized eligibility criteria

Type of criteria	Count
TS = prior, TIS = absent	13.6%
Criteria requiring confirmation by particular test	3.8%
Criteria with temporal constraint	9.2%
Criteria with value constraints	24.5%
Criteria containing some exception	9.6%

The detailed evaluation of precision of obtained results should be addressed in future work. It depends on precision of pattern detection algorithm, MetaMap, GATE semantic taggers and comparison algorithm.

4.2 Scenarios of Usage

The following scenarios show how the library of criteria could enhance the reuse of formalized criteria by trials designers. Modeling the content of eligibility criteria enabled browsing the library with the fine-grained queries, which correspond to the properties of patterns and instantiating concepts, and find criteria that:

1. Mention a specific concept e.g. 'Tamoxifen'
2. Mention a specific concept in a particular context. Following examples present criteria that mention Tamoxifen in various semantic contexts:

Context	Example of criteria related to Tamoxifen
TS= Planned event	Must be scheduled to receive adjuvant chemo-therapy with or without tamoxifen
TIS= Absence	No concurrent tamoxifen
ST= Mental or Behavioral Dysfunction	No serious toxicity (e.g. depression) thought to be due to tamoxifen
CT= Temporal constraint	At least 12 months since prior tamoxifen, raloxifene, or other antihormonal therapy

3. Mention some concept with a specific semantic type e.g.:

Semantic type	Example of criteria
Enzyme	Transaminases less than 3 times normal
Hormone	No adrenal corticosteroids
Laboratory procedure	Fasting blood glucose normal

4. Have specific domain e.g.:

Content domain	Example of criteria
Biologic therapy	No prior bone marrow transplantation
Cardiovascular	No history of deep vein thrombosis
Neurologic	No dementia or altered mental status

5. Are less restrictive than provided criterion e.g.:

Criterion	Possible relaxation
1. Creatinine < 1.2 mg/dL	Thresholds: 1.3, 1.8, 2.2, 2.5
2. At least 3 months since prior hormonal therapy	Thresholds: 1, 2, 3, 4 weeks
3. No prior endocrine therapy	No prior hormonal therapy for breast cancer
4. No prior malignancy	No other malignancy within the past 5 years except nonmelanomatous skin cancer or excised carcinoma in situ of the cervix

The first and second case represent examples of relaxations based on identifying less restrictive value thresholds. It is worth noting that because of using normalized representations of measurements, it was possible to compare number of months and weeks, as both thresholds were represented in seconds. Suggesting a threshold that was used by another medical expert should be more relevant than suggesting any arbitrary lower value. In the third case (No prior endocrine therapy), the potential relaxation was identified because of detecting ontology concepts, endocrine and hormonal therapy are synonyms, have the same UMLS identifier. The consequence of using this relaxation would be inclusion of patients that obtained such treatment for another purpose than breast cancer. The last example (No prior malignancy) represents a case of finding a relaxation based on both temporal constraint and stated exception. This alternative criterion considers eligible patients who had malignancy more than 5 years ago, or patients with such specific type of disease e.g. nonmelanomatous skin cancer. There is a significant need for providing meaningful suggestions, which can be illustrated by the fact, that searching for the subtypes of malignant disorder only in SNOMED CT, which is one of many ontologies covered by UMLS, returns 48 hits. Proposing those that were used in other eligibility criteria is a way of implicit incorporation of domain knowledge. However, the medical relevance of such suggestions will need to be verified by medical experts.

Apart from finding relevant criteria, the model enables to track their source, the trials where they are mentioned, and browse other criteria that they specify.

Presented methods for knowledge extraction from natural text could be possibly applied for other types of specialized language.

5 Related Work

There are several repositories that contain clinical trial data. The major one is ClinialTrials.gov, at the date of access contained 125301 trials. Its search engine

allows browsing the content by specifying trial metadata such as phase of a trial, interventions used etc. However, besides age and gender other eligibility criteria are not structured. Another source of clinical trial data is provided by the LinkedCT project [6], published according to the principles of Linked Data, enriched with the links to other sources. This repository has the same limitation; namely eligibility criteria are represented as free text.

Many studies have focused on the problem of formalization of eligibility criteria and clinical trial matching. There are several languages, which could be applied for expressing eligibility criteria e.g. Arden syntax [7], Gello [8], ERGO [9] and others. Weng et al [10] present the rich overview of existing options. However, no complete solution to the problem of automatic formalization of free text of criteria has been published. A considerable amount of work in that area is described in [11], where the authors describe their approach to semi-automatic transformation of free text of criteria into queries. It is based on manual preprocessing steps and further, automatic annotation of text with the elements of ERGO, which is a frame-based language. The authors describe how the results can be used to create the library of conditions, organized as a hierarchy of DL expressions, generated from ERGO annotations. They also note that creating such library could help creating criteria more clearly and uniformly. Because of the required manual steps the method cannot be directly reused.

Understanding free text of eligibility criteria could benefit from the information retrieval field, the overview of machine learning and knowledge based methods for relation extraction can be found in [12].

The general task of supporting design of clinical trials has not been broadly addressed in the literature. The system Design-a-trial [13] provides support for design of statistical measurements, i.e. suggesting minimal number of participants and kind of statistical test, ethical issues (e.g. choosing a drug with the least side effects) and preparing required documentation. It does not provide the support for designing eligibility criteria.

6 Conclusions and Future Work

This paper presents the study we conducted with the aim to enhance the reuse of structured eligibility criteria in order to support trial design.

We described our method for automatic formalization of eligibility criteria and the comparison of their restrictiveness. It is based on our pattern detection algorithm, and applies the semantic taggers from GATE, and MetaMap ontology annotator. Using our method we processed eligibility criteria from 300 clinical trials, and created a library of structured conditions. The library covers 18 % of encountered inclusion and exclusion criteria. It can be browsed with fine-grained queries thanks to the detailed modeling of criteria content. The supported scenarios of usage allow searching for eligibility criteria that mention specific data items in particular context, defined by various dimensions (temporal status, time independent status, specification type) and that are less restrictive than a given criterion. The method can be directly used for another trial set.

Possibly, similar strategy for knowledge extraction from natural text could be applied also for documents from other domains defined with specialized language.

The presented study needs additional research. The precision of the methods should be verified. The next challenge is to improve the scope of the library. The first obvious way is to increase the number of clinical trials used for populating. Another more interesting line is to improve the recall of the method for criteria formalization, to increase the variety of criteria that are covered. Additionally, we plan to assess the applicability of presented methods using patient data and feedback of clinical investigators about reusability of structured criteria, and medical relevance and ranking of provided suggestions. Although we created the bases, the empirical study should verify whether it can lead to increased enrollment of patient into clinical trials.

If the library of criteria is linked to a hospital database (EHR), another interesting issue 'trial feasibility' could be addressed using historical patient data. Namely, we could provide information about the consequence of modifying a given criterion in a certain way on the number of potentially eligible patients.

References

1. Cuggia, M., Besana, P., Glasspool, D.: Comparing semi-automatic systems for recruitment of patients to clinical trials. International Journal of Medical Informatics 80(6), 371–388 (2011)
2. Milian, K., ten Teije, A., Bucur, A., van Harmelen, F.: Patterns of Clinical Trial Eligibility Criteria. In: Riaño, D., ten Teije, A., Miksch, S. (eds.) KR4HC 2011. LNCS, vol. 6924, pp. 145–157. Springer, Heidelberg (2012)
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the ACL (2002)
4. Aronson, A.R.: Metamap: Mapping text to the umls metathesaurus. In: Proceedings AMIA Symposium (2001)
5. G. Units, Units (November 2006), `http://www.gnu.org/software/units/`
6. Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R.J., Wang, M.: Linkedct: A linked data space for clinical trials, CoRR abs/0908.0567
7. Wang, S., Ohno-Machado, L., Mar, P., Boxwala, A., Greenes, R.: Enhancing arden syntax for clinical trial eligibility criteria. In: Proceedings AMIA Symposium
8. Sordo, M., Ogunyemi, O., Boxwala, A.A., Greenes, R.A.: Gello: An object-oriented query and expression language for clinical decision support. In: Proceedings AMIA Symposium, vol. (5), p. 1012 (2003)
9. Tu, S., Peleg, M., Carini, S., Rubin, D., Sim, I.: Ergo: A templatebased expression language for encoding eligibility criteria. Tech. rep. (2009)
10. Weng, C., Tu, S.W., Sim, I., Richesson, R.: Formal representations of eligibility criteria: A literature review. Journal of Biomedical Informatics (2009)
11. Tu, S., Peleg, M., Carini, S., Bobak, M., Rubin, D., Sim, I.: A practical method for transforming free-text eligibility criteria into computable criteria
12. Cheng, X.-y., Chen, X.-h., Hua, J.: The overview of entity relation extraction methods. Intelligent Computing and Information Science 134, 749–754 (2011)
13. Nammuni, K., Pickering, C., Modgil, S., Montgomery, A., Hammond, P., Wyatt, J.C., Altman, D.G., Dunlop, R., Potts, H.W.W.: Design-a-trial: a rule-based decision support system for clinical trial design. Knowledge-Based Systems